Malaysian Journal of Economic Studies 51(2): 201-215, 2014 ISSN 1511-4554

Modelling Count Data with Excess Zeros: An Application to Health Care Utilisation Data

Shamzaeffa Samsudin*
Universiti Utara Malaysia

Peter G. Moffatt**
University of East Anglia

Abstract: This study is concerned with the estimation of microeconometric models of health care utilisation. The data set consists of 14,706 individuals from the General Household Survey for Great Britain, and the dependent variable is the number of General Practitioner (GP) consultations over a 2-week period. A clear feature of this count variable is excess zeros, and it is essential to incorporate this feature in the modelling strategy. Accordingly, in addition to standard Poisson and negative binomial models, zero-inflated, two-part, and latent class models were estimated. The zero-inflated negative binomial model (ZINB) proved, on the basis of several selection criteria, to be superior to other models for this dataset. As anticipated, health related variables had significant effects in determining health care utilisation while socioeconomic variables appeared to be less important, according to the results from the preferred model. Some effects differ quite markedly between the different models, underlining the importance of the type of process used to identify the best-fitting one.

Key words: Count data, excess zeros, health care utilisation, Poisson model, zero-inflated negative binomial
JEL classification: C50, I11, I18

1. Introduction

In survey data where the reference periods are short, zero occurrences of count events are inevitable. Respondents, for instance, may report zero utilisation when asking to report the number of doctor consultations in the last two weeks or three months before the interview. Zeros would be of two types here: users who have not utilised the services within the reference period, known as frequency zeros; and zeros reported by non-users. The latter type of zeros may be considered analogous to abstention in the consumption context. Clearly, a short reference period leads to a large number of frequency zeros. However, short reference periods also have the advantage of minimising recall-bias among respondents. Events that might be easier to recall, such as hospitalisation episodes, may have longer reference periods of perhaps six or twelve months.

* School of Economics, Finance and Banking, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia. Email: *shamzaeffa@uum.edu.my* (Corresponding author)

** School of Economics, University of East Anglia, Norwich, NR4 7TJ, UK. Email: *P.Moffatt@uea.ac.uk*

The data are obtained from the General Household Survey 2004-2005 produced by the Office for National Statistics-Social and Vital Statistics Division, UK. We would like to thank the editor and the anonymous referee for their insightful comments. All remaining flaws are the authors' responsibility.

Like unobserved heterogeneity, the presence of excess zeros gives rise to over-dispersion. Several estimation approaches are available for count data with excess zeros. These models include zero-inflated models (Ateca-Amestoy and Prieto-Rodriguez 2013; Gerdtham 1997; Sarma and Simpson 2006) which distinguish between frequency zeros and abstention, two-part or hurdle models (Deb and Trivedi 2002; Gerdtham 1997; Pohlmier and Ulrich 1995) and latent class (sometimes known as finite mixture) models (see Atella *et al.* 2004; Bago d'Uva,2006; Deb *et al.* 2006; Gerdtham and Trivedi 2001).

This study sets out to identify the best-fitting model for a particular count data set with excess zeros: the utilisation of general practitioners consultations (known as GP consultations henceforth) from the General Household Survey 2004/2005 for Great Britain. We estimated nine different models and used several model selection criteria to identify the best-fitting one. The results from the preferred model were then compared with those from the standard count data models.

The rest of the paper is organised as follows: In Section 2, we discuss the modelling strategies. Section 3 introduces the data and provides summary statistics. Section 4 reports and discusses the results. Section 5 concludes.

2. Empirical Specifications

A number of different approaches is taken in modelling demand for health care, and the choice depends on the nature of the data under analysis. This section outlines several models that have been used in the literature when modelling health care with a count dependent variable. The specifications of the models are based on Cameron and Trivedi (1986; 2006), Deb and Trivedi (1997; 2002) and Winkelmann and Zimmermann (1995). We begin with the standard count data models and then discuss several alternatives for the modelling of data that exhibit excess zeros. Throughout the discussions, y_i is used to represent the observed value of random variable *Y* (number of utilisations) for individual *i*, *i* = 1,...,N, N is the sample size.

2.1 Standard Poisson and Negative Binomial

The initial empirical model for health care demand, *Y*, is specified as below:

$$\mathrm{E}(Y = y_i|x_i) = \exp(x_i'\beta), i = 1...N, \quad (1)$$

where y_i is the realised demand for health care for individual *i* and x_i is a vector of characteristics of individual *i*, assumed to be exogenous, that determine y_i. Since the dependent variable is restricted to non-negative integer values, count data models are required. The most popular of these models are the Poisson and negative binomial (NB) models. As no assumption has been made to distinguish different decision processes between the contact decision and the frequency of utilisation, hurdle specifications are not called for at this stage. Besides, it is difficult to identify from the survey whether multiple utilisations are from the same episode of illness, which makes it difficult to differentiate the types of process. Suppose the number of occurrences for y_i given x_i, is Poisson distributed with density:

$$f(y_i|x_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \; y_i = 0,1,2...$$

with the consequence that

$$E(y_i|x_i) = \lambda_i = \exp(x_i'\beta) = V(y_i|x_i) \quad (2)$$

Equation (2) shows the equality of the conditional mean and conditional variance (equi-dispersion). Count data may turn out to be over-dispersed as a consequence of either unobserved heterogeneity, event dependence over time, or excess zeros. In any of these situations, the equi-dispersion assumption is violated, and the Poisson model fails.

Suppose, for every individual *i*, we introduce the random term that may be caused by specification error or unobserved heterogeneity, ε_i, into the conditional mean function of the Poisson model as follows:

$$E[y_i|x_i, \varepsilon_i] = \exp(x_i'\beta + \varepsilon_i); \; y_i = 0,1,2.....$$

$$= \lambda_i \upsilon_i; \; \lambda_i = \exp(x_i'\beta) \text{ and } \upsilon_i = \exp(\varepsilon_i)$$

Conditional on x_i, and with some algebraic manipulation, it can be shown that *Y* has a NB distribution with density function given by:

$$\Pr(y_i|x_i) = \frac{\Gamma(y_i + \psi_i)}{\Gamma(y_i + 1)\Gamma(\psi_i)} \left(\frac{\psi_i}{\lambda_i + \psi_i}\right)^{\psi_i} \left(\frac{\lambda_i}{\lambda_i + \psi_i}\right)^{y_i}, \quad y_i = 0,1,2,.... \quad (3)$$

where $\Gamma(.)$, is the *gamma* function, the index $\psi_i = (1/\alpha)\lambda_i^k$, $\alpha > 0$ is an over-dispersion parameter and *k* is a constant. The mean and variance functions are specified as:

$$E(y_i|x_i) = \lambda_i \quad \text{and} \quad V(y_i|x_i) = \lambda_i + \alpha\lambda_i^{2-k}$$

There are two variance functions depending on *k*. If we set *k* equal to 1, the variance becomes proportional to the mean (known as the NB1 model) while by setting *k* to 0, the variance becomes a quadratic function of the mean [known as the NB2 model (Cameron and Trivedi 1986)]. The model will simplify the Poisson if α equals 0. The likelihood ratio (LR) and Wald tests are used to test for over-dispersion of all equations by considering the quadratic variance function of the negative binomial model:

$$V(y_i | \lambda_i, \alpha) = \lambda_i + \alpha\lambda_i^2$$

The null hypothesis for the over-dispersion test is when α equals 0. The LR and Wald tests are used for this purpose. Since errors may be correlated within the household, cluster-robust standard errors are used in all models.

2.2 Zero-inflated Model

The zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models take into account distribution with excess zeros and attempt to resolve it by adding extra weight to the probability of a zero observation. In this model, individuals are split into two categories: non users and potential users, or according to Deb and Trivedi (1997), the 'not at risk' and 'at risk' population. These models may be interpreted as a combination of two processes: the first process generates zero observations; the second generates observations from either the Poisson or the negative binomial, allowing both non-zero and zero values to be observed. Zeros resulting from the second process would be from

individuals that have the potential to become health care users but have zero utilisation during the reference period.
Suppose we observe

$$\sim \begin{cases} 0 & \text{with probability } \varphi_i \\ g(y_i|x_i) & \text{with probability } 1-\varphi_i \end{cases}$$

The conditional probability of observing y_i is given by

$$\Pr(y_i|x_i,z_i)=\begin{cases} \varphi_i(z_i'\gamma)+\{1-\varphi_i(z_i'\gamma)g(0|x_i)\} & \text{if } y_i=0 \\ \{1-\varphi_i(z_i'\gamma)g(y_i|x_i)\} & \text{if } y_i>0 \end{cases} \quad (4)$$

The vector of zero-inflated covariates is z_i' and γ is the vector of zero inflated coefficients to be estimated. The term $\varphi_i(z_i'\gamma)$ can be modelled as either logit or probit while $g(y_i|x_i)$ is either a Poisson or negative binomial distribution. The logit function is used to model $\varphi_i(z_i'\gamma)$ in the analysis of this chapter.

2.3 Two-Part Model

To understand health care demand in the light of the two-part model (TPM) - also known as the hurdle model - we may divide the individual's decision process into two stages. The first stage is the decision whether to consume health care in a certain period of time; the second stage is when the health care provider, after the first contact, determines the next visit(s). This model has been associated with the principal-agent framework, in which the doctor acts as an agent for the patient (principal) and demands health care on behalf of the patient, on the basis that the doctor can be expected to know more than the patient about the type of health care needed. This model, however, assumes that the demand is determined by a single spell of illness. In a simple way (see Deb and Trivedi 2002; Sarma and Simpson 2006) for further analysis), we can show that the probability of these two distinct processes are given by

y_i

$\Pr(y_i = 0) = f_1(0)$

and

$$\Pr(y_i|y_i>0)=\frac{1-f_1(0)}{1-f_2(0)}f_2(y_i), \quad y_i>0 \quad (5)$$

It collapses to the standard model if $f_1(.)=f_2(.)$

The two parts of the model are estimated separately: the first as a binary model, that is, either probit or logit; and the second as a truncated count data model: either the truncated Poisson or truncated negative binomial. In some health care systems where a GP acts as a gatekeeper, the utilisation of health care services like outpatient and inpatient are jointly determined by the patient and the GP in the first stage. Hence, the decision to hospitalise or consult a doctor in the hospital cannot be interpreted as similar to a GP consultation.

A number of studies have utilised the two-part model. Pohlmier and Ulrich (1995) employed a negative binomial hurdle model to explain the demand for health care. They suggest that a two-part model is essential because of different decision processes whereby

the initial visit to the physician is determined by the individual while the frequency is decided by the physician. Jiménez-Martín *et al.* (2002) compared the two-part models with the latent class models in order to estimate demand for physician services of twelve countries in European Union.

By using the two model selection criteria of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), it was found that two-part models are more favoured for the specialist demand framework while latent class models are better in explaining demand for GPs. Two-part models have also been estimated by Mocan *et al.* (2004) and Sarma and Simpson (2006).

2.4 Latent Class Model

The latent class model (LCM) is another mixture model that accommodates the problem of excess zeros. It allows for individual heterogeneity by dividing the population into several latent classes based on unobserved criteria, for example, an individual's long term health status (Deb and Trivedi 1997; 2002). Unlike the two-part model, which is also a mixture model, the LCM is believed to be more flexible as it does not restrict itself to the distinction between zero and positive values.

Suppose the population is divided into C latent classes in proportion $\pi_1, \pi_2, \ldots \pi_c$. The density function can be specified as:

$$f(y_i|\Theta) = \sum_{j=1}^{c-1} \pi_j f_j(y_i|\theta_j) + \pi_c f_c(y_i|\theta_c), \qquad i = 1,2.....n, \quad j = 1....,C \tag{6}$$

where $\pi_1 \geq \pi_2 ..... \geq \pi_c = \left(1 - \sum_{j=1}^{c-1} \pi_j\right)$ are the mixing proportions which are to be estimated[1] along with other parameters from all components, $\theta_1, \theta_2 \ldots \theta_c = \Theta$.

The component density for the finite-mixture Poisson and negative binomial is similar to the standard density function of those models but varies across components. The mean and variance functions for the finite-mixture Poisson are given by:

$$\mathrm{E}(y_i|x_i) = \mathrm{V}(y_i|x_i) = \sum_{j=1}^{c} \pi_j \lambda_{ji} = \overline{\lambda}$$

while for the finite-mixture negative binomial, it is:

$$\mathrm{E}(y_i|x_i) = \sum_{j=1}^{c} \pi_j \lambda_{ji} = \overline{\lambda} \text{ and } \mathrm{V}(y_i|x_i) = \sum_{j=1}^{c} \pi_j \lambda_{ji}^2 \left[1 + \alpha_j \lambda_{ji}^{-\lambda}\right] + \overline{\lambda}_i - \overline{\lambda}_i^2$$

Using data from the Canadian National Population and Health Survey, Sarma and Simpson (2006) compared the hurdle model with the LCM. They found that both the AIC and BIC suggest that LCM is preferred to the hurdle model for doctor and GP visits. Using the data from the RAND Health Insurance Experiment, Deb and Trivedi (2002) also found that the AIC and BIC favour the LCM over TPM. However, the LCM seems to be motivated by statistical convenience rather than any sort of theoretical rationale, of a two-part

[1] The latent class models in this study have been fitted using Stata's user-written command in Stata 10 by Partha Deb, Hunter College and The Graduate Center, City University of New York.

process in the health care utilisation decision, of the type that motivates the two-part model.

3. Data and Summary Statistics

For this empirical analysis, data from the General Household Survey (GHS) 2004/2005 for Great Britain were used. Since the main objective of this study is to analyse the performance of different models, the age of the dataset is of minor importance. The GHS was used because it contains rich information on the frequency of utilisation of various types of medical care, as well as other important variables that might influence the utilisation level. This information allows the use of count data models in analysing the uses of medical care. The 2004/2005 survey covered 8,700 households consisting of 20,421 individuals of all ages. All adults aged 16 or over were interviewed while proxies were used to answer on behalf of children. Two-stage sampling was used. In the first stage, Primary Sampling Units (PSU) were selected from postcode sectors; in the second stage, Secondary Sampling Units (SSU), were selected as addresses within the sampled sectors. All individuals or proxies within selected households were interviewed.

Due to missing data in some variables of interest, only 14,706 observations were left for data analysis which represented 72% (the *reduced sample* henceforth) of the original sample. All observations that had at least one missing value in variables used were deleted. It turned out that the maximum age for the reduced sample was only 69 compared to 99 in the original sample. This variable was checked and it was confirmed that all observations aged above 69 were dropped from the estimation sample due to missing values for education level.

In this case, all analyses and discussions were confined to members of the population aged 0 to 69 only. Explanatory variables were selected based on previous literature on health care demand. This selection has been narrowed down from a large set of variables to the set reported in Table 1. Table 1 describes the variables used in the analysis together with summary statistics (see Appendix 1 for details). We also present the summary statistics based on the type of users in Appendix 2.

4. Results and Discussion

Figure 1 shows the frequency distribution (%) of doctor consultations. This histogram clearly reveals the existence of excess zeros, suggesting that standard count data models may not be sufficient.

Nine different models were estimated and compared (Table 2). They are: the standard Poisson and negative binomial, zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), two-part probit-truncated Poisson (TPP), two-part probit-truncated negative binomial (TPNB), latent class Poisson-two components (LCP-2), latent class negative binomial- two components (LCNB-2) and latent class negative binomial-three components (LCNB-3). In negative binomial regressions, all variances were specified to have a quadratic function of the mean (NB2, as explained in Section 2) as some models using NB1 specification failed to converge. The LCNB-3 model also failed to converge. The standard errors reported here are based on clustered sandwich estimators that allow for the correlation between individuals in the same household.

Table 1. Dependent and explanatory variables

Variables	Definitions	Mean	Sd.	Min	Max
Dependent GP	Number of GP consultations for the past 2 weeks	0.184	0.515	0	7
Explanatory					
I. Personal characteristics					
AGE	Age in years	34.317	19.664	0	69
AGESQ	Square of age in years/100	15.643	13.720	0	47.6
MALE	1 if gender is male, 0 if female	-	-	0	1
Marital status					
SINGLE	1 if single, widowed, divorce, separated, 0 otherwise	-	-	0	1
COHAB	1 if cohabitate, 0 otherwise	-	-	0	1
MARRIED	1 if married, 0 otherwise	-	-	0	1
II.Health status and health related variables					
GOODHLTH	1 if assessed health state is good, 0 poor	-	-	0	1
LIMITACT	Number of days with activities prevented	0.940	3.100	0	14
LONG_ILL	Number of longstanding illness	0.413	0.808	0	6
III. Socio-economic status					
Education					
HIGH_EDU	1 if has higher qualification, 0 otherwise	-	-	0	1
OTHER_EDU	1 if has other qualification, 0 otherwise	-	-	0	1
NO_EDU	1 if has no qualification, 0 otherwise	-	-	0	1
INCOME	Log of equivalised household income	5.054	1.089	0	9.6
Country					
ENGLAND	1 if live in England, 0 otherwise	-	-	0	1
WALES	1 if live in Wales, 0 otherwise	-	-	0	1
SCOTLAND	1 if live in Scotland, 0 otherwise	-	-	0	1
GPPOP	Number of GP per thousand population	0.640	0.048	0.58	0.74

Notes: 1. Variables in *ITALICS* are the reference variables; 2. Marital status of children has been recoded according to HRP marital status. Cohabitate is living together without being legally married; 3. Health status is a self-perceived health state in the last 12 months before the interview. Health status 'fairly good' and 'not good' were combined as 'poor' status; 4. Higher qualification includes higher degree, first degree, teaching qualification, other higher qualification, and nursing qualification. Other qualifications include GCE A level in two or more subjects, GCE A level in one subject, GCSE/O LEVEL, standard grades, GCSE/O LEVEL, GCSE below grade 1, GCSE below grade C, apprenticeship and other vocational, professional or foreign qualifications.

The model selection processes are summarised in Figure 2. First, selection is made between nested models. These selections are between the Poisson and negative binomial; between ZIP and ZINB; between TPP and TPNB; and between LCP-2 and LCNB-2.

Selections were made using the likelihood ratio (LR) test in each case with null hypothesis that the over-dispersion parameter α equals 0. In all cases, the LR tests indicate that models with NB densities are strongly favoured over the corresponding Poisson models. For this reason, further comparison is restricted to models that are based on negative binomial densities; these are standard NB, ZINB, TPNB and LCNB-2.

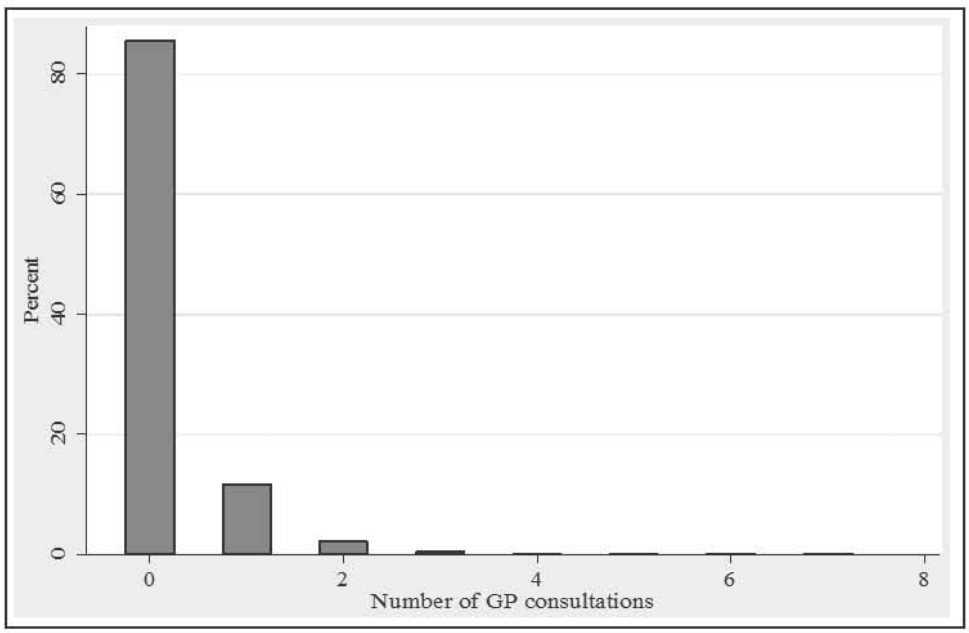

Figure 1. *Frequency distribution of GP consultations*

To select the best model, three model selection criteria were used: the log likelihood value (LL), the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).[2] For each criterion, values were compared across models. If LL is used as a criterion, the model with the highest LL is favoured. If AIC or BIC is used, the model with the lowest value is favoured. The comparisons made using the bottom row of Figure 2 shows a unanimous pattern of selection that favours the ZINB model over other models. Note that ZINB is also preferred to the standard negative binomial specification (NB2) and this is established using the Vuong test.[3] In ZINB, the split between potential users and non-users is set to be determined by all covariates within the model. For comparison, results from other extended models are also presented in Table 2. The interpretation of the results depends on the model selected.

In our preferred model (ZINB), health care users can be divided into two categories: potential users and non-users. From the first column of ZINB results, we see that variables like MALE, GOODHLTH and LIMITACT are significantly determining consultations among potential users with directions similar to those from other competing models. People with a good health utilise less health care than people with poor status while an increase in the number of activities contributed to more utilisation. The number of longstanding illnesses (LONG_ILL) and education level, however, show no significant effect for potential

[2] AIC formula=-2log(L)+2K and BIC=-2log(L)+Klog(N); where K is the number of parameters and N is the number of observations. Log(L) is the value of maximised log-likelihood.

[3] The Vuong test is used because according to Greene (1994), the two models are not nested.

Table 2. Negative binomial, zero-inflated, two-part and latent class estimates for GP consultations; N=14,706

	Negative binomial		Zero-inflated negative binomial (ZINB)				Two-part negative binomial (TPNB)				Latent Class NB-2			
					Inflate (non-user)		Zero truncated NB		Probit		Component 1		Component 2	
	coef	s.e	coef	s.e	coef	s.e	coef	s.e	coef	s.e	coef	s.e	coef	s.e
AGE	-0.008	0.005	-0.005	0.006	0.007	0.008	0.021*	0.012	-0.008**	0.003	-0.019**	0.016	0.005	0.022
AGESQ	0.006	0.007	-0.007	0.008	-0.027**	0.012	-0.043**	0.018	0.009**	0.004	0.019*	0.023	-0.010	0.032
MALE	-0.305***	0.045	-0.185***	0.064	0.233***	0.089	-0.050	0.112	-0.199***	0.027	-0.465***	0.133	-0.115	0.128
COHAB	0.053	0.084	0.117	0.105	0.100	0.140	0.198	0.181	0.011	0.051	-0.114	0.197	0.261	0.225
MARRIED	0.020	0.052	0.061	0.068	0.083	0.097	0.056	0.122	0.008	0.032	-0.016	0.137	0.070	0.188
GOODHLTH	-0.859***	0.053	-0.514***	0.087	0.319***	0.115	-0.614***	0.136	-0.478***	0.032	-1.130***	0.242	-0.558***	0.200
LIMITACT	0.097***	0.005	0.059***	0.007	-5.304***	0.821	0.073***	0.009	0.070***	0.004	0.089***	0.016	0.106***	0.017
LONG_ILL	0.131***	0.024	0.019	0.029	-0.560***	0.116	0.058	0.052	0.106***	0.018	0.131***	0.044	0.131*	0.071
INCOME	-0.015	0.023	0.005	0.033	0.052	0.058	-0.004	0.055	-0.011	0.014	-0.025	0.049	-0.005	0.070
OTHER_EDU	-0.111**	0.056	-0.050	0.076	0.064	0.100	-0.041	0.145	-0.069**	0.033	-0.040	0.115	-0.195	0.158
NO_EDU	-0.146**	0.068	0.017	0.088	0.228*	0.131	0.088	0.168	-0.104**	0.042	-0.133	0.166	-0.147	0.222
WALES	-0.065	0.126	0.062	0.152	0.217	0.178	0.044	0.287	-0.060	0.071	-0.025	0.342	-0.122	0.467
SCOTLAND	-0.064	0.112	-0.102	0.147	-0.099	0.223	-0.095	0.254	-0.032	0.071	-0.221	0.239	0.142	0.316
GPPOP	0.822	0.704	0.842	0.959	-0.308	1.446	1.439	1.666	0.442	0.438	2.010	1.571	-0.798	2.130
CONSTANT	-1.551***	0.465	-1.131*	0.626	-0.038	0.955	-2.977***	1.131	-0.885***	0.291	-2.026**	0.903	-0.847	1.258
α	0.857	0.100			0.297	0.247	2.668	2.139			0.000	0.000	2.086	1.110
π											0.643	0.110	0.357	0.110
LogL	-6842.7379		-6693.26				-6787.46				-5457.188			
Vuong			8.31											

Notes: The symbols ***, **, and * denote 1, 5 and 10% level of significance, respectively.

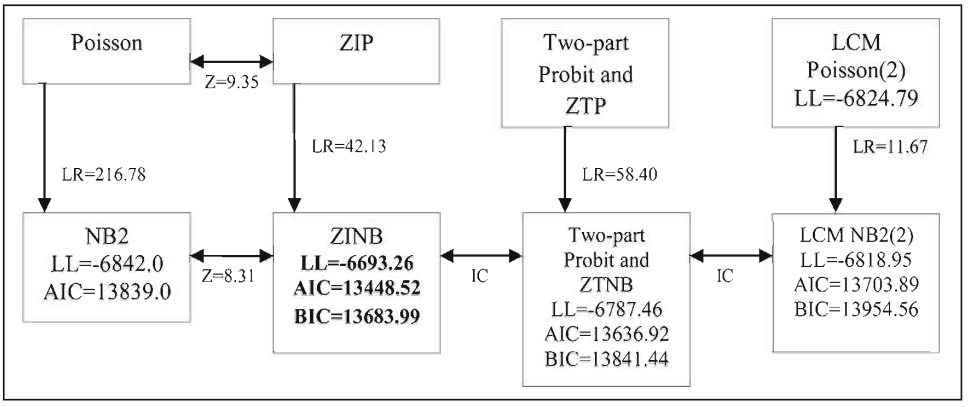

Figure 2. *The selection process of the best model*

Notes: 1. Single-pointed arrows represent the nested tests while double-pointed arrows indicate the non-nested tests; 2. The Likelihood ratio test statistics (LR) are used to select between the nested models while Z statistics are from Vuong tests; the AIC and BIC are used for the non-nested models; 3 All LR tests prefer the NB2 specifications over the Poisson while Vuong tests prefer the inflated models (ZIP and ZINB); 4. Values in bold indicate the best value according to LL, AIC or BIC.

users in which without a separate classification as in NB model, these variables have significant effects.

The second column of ZINB results tells us what factors determine the probability of being a non-user. Hence we expect signs in this equation to be opposite from signs in the first equation. MALE, GOODHLTH and LONG_ILL are among variables that have significant effects on the probability of being a non-user. Individuals with good health status, male, or who have had no education are more likely to be non-users of health care. The effect of education for non-users, however, is not consistent with the theoretical role of education in the Grossman theory of health demand (Grossman 1972) which suggests that the efficiency of producing health stock depends on other forms of human capital, which include education. According to Grossman, people with education are believed to have higher productivity in producing better health, and thus require less health care. Consistent with other studies, males utilise health care less frequently (Atella *et al*. 2004; Gerdtham 1997; Hunt-McCool *et al*. 1995).

These results reflect the high level of equity of health care utilisation in the UK as it should highly depend on health status rather than other variables. Nevertheless, understanding the influence of significant variables on health care utilisation is important in formulating a health policy in order to achieve its objectives. In improving the health status of the population that in turn may also control health care use in the future, policy analysts may continually consider alternatives on how to improve health status through health, social or education policy. Policies on health promotion that include health education, prevention and protection should be enhanced and strengthened as it is well known that health knowledge would increase the ability of individuals to enjoy good health in the long run. It is also established that participation in health education

would increase health knowledge of individuals while formal education improves the efficiency of producing the knowledge.

5. Conclusion

This study attempts to compare the performance of competing micro-econometric models for count data that contain excess zeros. Model selection criteria suggest that standard count models are not sufficient for modelling data of this type. Effects may vary across models, and it is therefore important to select the best-fitting model for better interpretation. The zero-inflated negative binomial (ZINB) model is preferred in modelling the utilisation of doctor consultations in this study. In the zero-inflated model, more weight is given to the probability of zero observation and effects are initially divided into two different classes. As found in other studies, ZINB is proven to be superior for modelling services with lower utilisation frequencies.

The performance of two-part specifications, which is believed to support the principal-agent approach, does not appear to be the best-fitting model for the dataset used in this study. The sign of effects in the GP equation are slightly different between standard and extended models but the associated standard errors (cluster-robust standard error) of the NB model are smaller than those of the zero-inflated model. From the ZINB model, there is statistical evidence that all health related variables - self-assessed health (GOODHLTH), number of days with activities prevented (LIMITACT) and number of longstanding illness (LONG_ILL) - are significant in determining GP consultations with the expected direction. Age, socio-economic variables (except for education), country and GP densities, on the other hand, do not have a significant influence on doctor consultations.

Although the data used in this study is within the context of the UK, a similar analysis could be applied to any dataset with excess-zero problems.

References

Ateca-Amestoy, V. and J. Prieto-Rodriguez. 2013. Forecasting accuracy of behavioural models for participation in the arts. *European Journal of Operational Research* **229(1)**: 124-131.

Atella, V., F. Brindisi, P. Deb and F.C. Rosati. 2004. Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach. *Health Economics* **13(7)**: 657-668.

Bago d'Uva, T. 2006. Latent class models for utilisation of health care. *Health Economics*, **15(4)**: 329-343.

Cameron, A.C. and P.K. Trivedi. 1986. Econometric models based on count data: Comparisons and Application of some estimators and tests. *Journal of Applied Econometrics* **1(1)**: 29-54.

Cameron, A.C. and P.K. Trivedi. 2006. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Deb, P., C.H. Li, P.K. Trivedi and D.M. Zimmer. 2006. The effect of managed care on use of health care services: Results from two contemporaneous household surveys. *Health Economics* **15(7)**: 743-760.

Deb, P. and P.K. Trivedi. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* **12(3)**: 313-336.

Deb, P. and P.K. Trivedi. 2002. The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics* **21(4)**: 601-625.

Gerdtham, U.G. 1997. Equity in health care utilisation: Further tests based on hurdle models and Swedish micro data. *Health Economics* **6(3)**: 303-319.

Gerdtham, U.G. and P.K. Trivedi. 2001. Equity in Swedish health care reconsidered: New results based on the finite mixture model. *Health Economics* **10(6)**: 565-572.

Greene, W.H. 1994. Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Models. Working Paper No. 94-10. New York: Stern School of Business, New York University.

Grossman, M. 1972. On the concept of health capital and demand for health. *The Journal of Political Economy* ***80*****(2)**: 223-255.

Hunt-McCool, J., B.F. Kiker and Y.C. Ng. 1995. Gender and the demand for medical-care. *Applied Economics* ***27*****(6)**: 483-495.

Jiménez-Martín, S., J.M. Labeaga and M. Martinez-Granado. 2002. Latent class versus two-part models in the demand for physician services across the European Union. *Health Economics,* **11(4)**: 301-321.

Mocan, H.N., E. Tekin and J.S. Zax. 2004. The demand for medical care in urban China. *World Development* **32(2)**: 289-304.

Pohlmeier, W. and V. Ulrich. 1995. An econometric-model of the 2-part decision-making process in the demand for health-care. *Journal of Human Resources* **30(2)**: 339-361.

Sarma, S. and W. Simpson. 2006. A micro-econometric analysis of Canadian health care utilization. *Health Economics* **15(3)**: 219-239.

Winkelmann, R. and K.F. Zimmermann. 1995. Recent developments in count data modelling: theory and application. *Journal of Economic Surveys* **9(1)**: 1-24.

Appendix 1

Variables	Questions/Information from GHS
GP	*During the 2 weeks ending yesterday, apart from any visit to a hospital, did you talk to a doctor for any reason at all, either in person or by telephone? If yes, how many times did you talk to a doctor in these 2 weeks.* Generate **GP** that includes 0 count (no consultation).
AGE	Age in years. Age squared (AGESQ) is the square of age divided by 100
MALE	male..1 female..2 Recode male=1, female=0 ; Rename male as MALE
SINGLE COHAB MARRIED	de facto marital status Married ……1 Cohabiting…… 2 Single ……3 Widowed ……..4 Divorced……5 Separated ……..6 Same sex couple…7 Marital status for children is based on marital status of the Household Reference Person (HRP) Recode Widowed, Divorced, Separated, Same sex couple=0; cohabiting=1; married=2. Rename dummies as SINGLE (single), COHAB (cohabiting), and MARRIED (married)
GOODHLTH	*Over the last twelve month would you say your health has on the whole been good, fairly good, or not good?* Self-Perceived General Health Good……..1 Fairly good…2 Not Good……3 Recode Good=1; Fairly Good plus Not Good=0 Rename Good as GOODHLTH, and Fairly Good plus Not Good as POOR
LIMITACT	*During the last two weeks, did you have to cut down on any of the things you usually do (about the house/at work or in your free time) because of illness or injury? If yes, how many days in all during 2 weeks, including Saturday and Sundays.* Generate LIMITACT that includes 0 count (no activities cut down).
LONG_ILL	*Do you have any longstanding illnesses or infirmity? By longstanding illness I mean anything that has troubled you over a period of time or that is likely to affect you over a period of time?If yes, how many?* Generate LONG_ILL that includes 0 count (no illness or injury)
HIGH_EDU OTHER_EDU NO_EDU	Higher qualification includes higher degree, first degree, teaching qualification, other higher qualification, and nursing qualification Other qualifications include GCE A level in two or more subjects, GCE A level in one subject, GCSE/O LEVEL, standard grades, GCSE/O LEVEL, GCSE below grade 1, GCSE below grade c, apprenticeship and other

Continued next page

Appendix 1. Continued from previous page

	vocational, professional or foreign qualifications. Education level for children is recoded based on education level of HRP. Rename dummies as HIGH_EDU (Higher qualification), OTHER_EDU (Other qualification), NO_EDU (No qualification)
INCOME	Net weekly equivalised household income in pence (nthheq) that includes all type of earnings, benefits, pension, dividends, interest and other regular payments, after deductions, received by all adults in the household. Generate INCOME= log(1+(nthheq/100))
ENGLAND WALES SCOTLAND	England …1 Wales……2 Scotland…3 Rename dummies ENGLAND (England), WALES (Wales), SCOTLAND (Scotland),
GPPOP	Ratio GP per thousand populations (GPPOP) Population number is based on the Government Office Region (GOR). The number of General Practitioners by GOR is retrieved from Regional Trends 2006 edition.

Appendix 2. Summary statistics of dependent and explanatory variables for users and non-users

	Users (n=2,125)				Non-users (n=12,581)			
Variables	Mean	Sd.	Min	Max	Mean	Sd.	Min	Max
Dependent GP	1.274	0.670	1	7	0	0	0	0
Explanatory								
I. Personal characteristics								
AGE	37.220	20.145	0	69	33.826	19.542	0	69
AGESQ	17.909	14.342	0	47.61	15.261	13.577	0	47.61
MALE	-	-	0	1	-	-	0	1
Marital status								
SINGLE								
COHAB	-	-	0	1	-	-	0	1
MARRIED	-	-	0	1	-	-	0	1
II.Health status and health related variables								
GOODHLTH	-	-	0	1	-	-	0	1
LIMITACT	2.960	5.019	0	14	0.600	2.486	0	14
LONG_ILL	0.803	1.105	0	6	0.347	0.726	0	6
III.Socioeconomic status								
Education								
HIGH_EDU								
OTHER_EDU	-	-	0	1	-	-	0	1
NO_EDU	-	-	0	1	-	-	0	1
INCOME	4.965	1.110	0	9.607	5.069	1.085	0	9.607
Country								
ENGLAND								
WALES	-	-	0	1	-	-	0	1
SCOTLAND	-	-	0	1	-	-	0	1
GPPOP	0.641	0.048	0.583	0.745	0.640	0.048	0.583	0.745

Variables in *ITALICS* are the reference variables.